ehp ENVIRONMENTAL HEALTH PERSPECTIVES

http://www.ehponline.org

A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations

Lydiane Agier, Lützen Portengen, Marc Chadeau-Hyam, Xavier Basagaña, Lise Giorgis-Allemand, Valérie Siroux, Oliver Robinson, Jelle Vlaanderen, Juan R. González, Mark J. Nieuwenhuijsen, Paolo Vineis, Martine Vrijheid, Rémy Slama, and Roel Vermeulen

http://dx.doi.org/10.1289/EHP172

Received: 26 May 2015
Revised: 12 January 2016
Accepted: 28 April 2016
Published: 24 May 2016

NIH National Institute of Environmental Health Sciences

Environ Health Perspect DOI: 10.1289/EHP172
Advance Publication: Not Copyedited

A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations

Lydiane Agier*[,1], Lützen Portengen*[,2], Marc Chadeau-Hyam[3], Xavier Basagaña[4,5,6], Lise Giorgis-Allemand[1], Valérie Siroux[1], Oliver Robinson[4,5,6], Jelle Vlaanderen[2], Juan R. González[4,5,6], Mark J. Nieuwenhuijsen[4,5,6], Paolo Vineis[3], Martine Vrijheid[4,5,6], Rémy Slama¶[,1], and Roel Vermeulen¶[,2,3]

[1]Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Inserm and Univ. Grenoble-Alpes, U823 Joint Research Center, Grenoble, France; [2]Institute for Risk Assessment Sciences, Utrecht University, Utrecht, the Netherlands; [3]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, Norfolk Place, London, W2 1PG, United Kingdom; [4]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [5]Universitat Pompeu Fabra (UPF), Barcelona, Spain; [6] CIBER Epidemiología y Salud Pública (CIBERESP), Spain.
*Joint first authorship. ¶Joint last authorship.

Address for correspondence Lydiane Agier, Institut Albert Bonniot, CRI INSERM/UJF U82, Rond-point de la Chantourne, 38700 La Tronche, France. Telephone: (0033) 4 76 54 94 00. E-mail: lydiane.agier@ujf-grenoble.fr

Running title: Regression-based methods for exposome studies

Acknowledgment: We acknowledge the input of HELIX - Exposomics statistical working group, in particular all participants to the meetings where this study was discussed. This work was supported by the European Community's Seventh Framework Programme FP7/2007-2013 (grants 308333, HELIX to Martine Vrijheid and 308610, EXPOsOMICS to Paolo Vineis).

Competing financial interest: The authors declare no conflict of interest.

Abstract

Background: The exposome constitutes a promising framework to better understand the effect of environmental exposures on health by explicitly considering multiple testing and avoiding selective reporting. However, exposome studies are challenged by the simultaneous consideration of many correlated exposures.

Objectives: We compared the performances of linear regression-based statistical methods in assessing exposome-health associations.

Methods: In a simulation study, we generated 237 exposure covariates with a realistic correlation structure, and a health outcome linearly related to 0 to 25 of these covariates. Statistical methods were compared primarily in terms of false discovery proportion (FDP) and sensitivity.

Results: On average over all simulation settings, the elastic net and sparse partial least-squares regression showed a sensitivity of 76% and a FDP of 44%; Graphical Unit Evolutionary Stochastic Search (GUESS) and the deletion/substitution/addition (DSA) algorithm a sensitivity of 80% and a FDP of 33%. The environment-wide association study (EWAS) underperformed these methods in terms of FDP (average FDP, 86%), despite a higher sensitivity. Performances decreased considerably when assuming an exposome exposure matrix with high levels of correlation between covariates.

Conclusions: Correlation between exposures is a challenge for exposome research, and the statistical methods investigated in this study are limited in their ability to efficiently differentiate true predictors from correlated covariates in a realistic exposome context. While GUESS and DSA provided a marginally better balance between sensitivity and FDP, they did not outperform the other multivariate methods across all scenarios and properties examined, and computational complexity and flexibility should also be considered when choosing between these methods.

INTRODUCTION

Environmental factors comprise a wide range of physical, chemical, biological and sociological stressors. As exemplified in twin- and migrant-studies, the environment may explain a relatively large fraction of the variation in the risk of many chronic diseases or continuous health traits (Rappaport et al. 2014; Willett 2002). Until now, studies in environmental epidemiology typically assessed the link between environmental exposures and health using approaches considering each environmental exposure separately, and therefore provided only a fragmented view of environment and health associations (Buck Louis et al. 2013; Rappaport 2011; Vrijheid et al. 2014) (see (Greenland 1994; Lenters et al. 2014) for exceptions). Results from these approaches suffer from possible confounding due to (ignored) co-exposures, selective reporting, and publication bias (Patel and Ioannidis 2014; Slama and Vrijheid 2015). The exposome concept, as originally defined by Wild (2005), comprises the totality of environmental exposures from the prenatal period onwards, and argues for a holistic consideration of all exposures simultaneously (Wild 2012).

Most previous studies relating the exposome to health relied on the Environment-Wide Association Study (EWAS, the association between each single exposure factor and the outcome being estimated separately) (Patel et al. 2010), sometimes followed by a multiple regression step that includes the selected predictors (Patel et al. 2013). Several multivariate regression-based statistical methods are now well established and allow accounting for a potential joint action of multiple exposures on health (Chadeau-Hyam et al. 2013). Sparse Partial Least Square (sPLS) (Chun and Keleş 2010) for instance has recently been used in a study of male fecundity (Lenters et al. 2014), while Elastic Net (ENET) (Zou and Hastie 2005) was used to link multiple environmental contaminants to birth weight (Lenters et al. 2015). To our knowledge, in the context of exposome research, no other multiple regression statistical method has yet been applied.

The statistical performances of these established statistical methods in an exposome context remain to be systematically assessed. In a recent simulation study (Sun et al. 2013), several multiple regression approaches were investigated for a limited number of exposures (N≤20), that were, at most, moderately correlated (Pearson correlation lower than 0.57). However in (future) exposome studies, many more covariates will likely be considered, and stronger correlations (typically greater than 0.6) are routinely observed in large exposome datasets, such as NHANES (Patel et al. 2010, Patel et al. 2013, Patel and Ioannidis 2014). We therefore extended the work by Sun et al. to a realistic exposome context and aimed to compare statistical performances of linear regression-based statistical methods for future exposome studies.

We generated exposure data using an empirical correlation structure between a large number of exposure covariates (i.e. 237), and assumed that 0 to 25 of these exposures linearly influenced a continuous health outcome without effect measure modification (i.e. interaction). The statistical methods compared included (i) the EWAS approach; (ii) EWAS followed by a multiple regression step including the identified hits; (iii) ENET, a penalized regression method; (iv) sPLS regression, a supervised dimension reduction method; (v) the Graphical Unit Evolutionary Stochastic Search (GUESS) algorithm, a computationally optimized Bayesian variable selection method (Bottolo et al. 2013), and (vi) the deletion/substitution/addition (DSA) sequential algorithm (Sinisi and van der Laan 2004). Statistical performances of selected approaches were systematically compared on the basis of six established criteria and two modified criteria, in order to evaluate both variable selection and point estimation. We additionally investigated the sensitivity of the statistical performances of the methods with respect to modifications of the empirical correlation structure used to generate the exposures.

METHODS

Our simulation model relied on generating a matrix of exposure variables X for a fictitious population. From this matrix, we generated the health outcome Y according to a linear regression model; seven scenarios were defined on the basis of the number of true predictors. We assessed the association between each simulated X and Y using a preselected set of statistical methods, whose performances were assessed for each scenario and compared using the metrics detailed below. For each scenario, we simulated 100 independent datasets.

Generation of the exposome

In order to generate exposure variables with a realistic correlation structure, we relied on the existing INMA (INfancia y Medio Ambiente) mother-child cohort (Guxens et al. 2012), in which a total of 237 environmental factors have been assessed in mothers during pregnancy through questionnaires, geospatial modeling and biological monitoring. From the matrix of all pairwise correlations, we computed the closest positive definite matrix (Higham 2002), and used this estimate as our benchmark correlation matrix Σ (Figure S1). We used Σ to generate X, the exposome of a virtual study population of 1200 subjects (size of the study population of an ongoing European exposome project comprising the INMA cohort (Vrijheid et al. 2014)) from a mean-centered multivariate normal distribution: $X \sim N(0, \Sigma)$. As the cohort data contained 5 binary variables (the others being continuous), we have dichotomized these 5 variables in our simulated datasets so as to replicate the proportion of positive responses observed in the original data.

Health outcome generation

The health outcome Y was generated as a function of the exposome according to

$$Y = \sum_{i=1}^{237} \beta_i X_i + \epsilon\,, \epsilon \sim N(0, \sigma^2)$$

Regression coefficients β_i were all set to 0 except for the k randomly selected variables that were assumed to be causally related to the outcome (hereafter referred to as "true predictors"), for which $\beta_i = 1$. We considered seven scenarios, each defined by a different number of true predictors: k=0,1,2,3,5,10,25. The residual variance σ^2 was defined such that the proportion of variance explained by the true predictors (R^2) equaled $3\% \times k$. With this constraint, within a given scenario the signal to noise ratio was the same in all simulations; and the power to select a true predictor in unadjusted analyses with uncorrelated true predictors was constant across scenarios (see Supplemental Material S1).

Seven versions of these scenarios were generated. Set 1 corresponds to the scenarios defined above. Sets 2 and 3 aimed to assess the impact of the correlation level amongst true predictors which could raise identifiability issues. These scenarios differed from set 1 by ensuring that correlation among all true predictors was in absolute value <0.2 for set 2, and >0.5 for set 3. Sets 4 and 5 aimed to assess the impact of the correlation structure of the whole exposome on the performances of the statistical methods; the scenarios differed from set 1 by not generating X from Σ but for set 4 from the correlation matrix Σ^{-}obtained by dividing the coefficients of Σ by two except on the diagonal; and for set 5 from Σ^{+} obtained by multiplying the coefficients of Σ by two, upper-bounding coefficients by 1 and computing the closest semi-definite matrix. Set 6 investigated deviating from the assumption of normally distributed exposures (i.e. including potentially skewed distributions and outliers) by generating scenarios similarly to set 1 except with exposure data obtained by bootstrapping the actual environmental data from the INMA cohort. Finally, set 7 investigated the methods' robustness to unequal effect sizes by generating scenarios similarly to set 1 except with effect sizes (i.e. β_i) for true predictors drawn from a uniform distribution in [0.5,1.5].

In all scenarios, the health outcome was generated as described above; for a given number of true predictors, the proportion of variance explained by the true predictors was therefore the same across all seven sets of scenarios. .

Statistical methods to estimate the exposome-health association

To estimate the association between Y and X, we used six linear regression-based statistical methods.

Environment-wide association study

The EWAS (Patel et al. 2010) relies on linear regression models fitted independently for each covariate. Statistical significance of the association between predictors and the response is assessed on the related 2-sided p-values after a correction for multiple comparisons was applied. As a benchmark, we considered the widely used Benjamini and Yekutieli (2001) correction to control the false discovery rate (FDR) at a desired level (here 5%). Additionally, covariates declared significant in the EWAS were included in a multiple linear regression model and retained if their 2-sided p-value was below 5% (Tzoulaki et al. 2012). This two-step approach is further referred to as EWAS-Multiple Linear Regression (EWAS-MLR).

As sensitivity analyses, we tested several procedures to correct for multiple hypothesis testing: a permutation-based approach (Patel et al. 2010), the Benjamini and Hochberg (1995) procedure and the Bonferroni (1936) correction. We also tested the EWAS method without applying a correction for multiple comparison as a way to illustrate what would happen if independent studies were applied for each exposure covariate separately.

Elastic net

The ENET (Zou and Hastie 2005) is a penalized regression model relying on a generalized linear framework, and uses a weighted mixture of the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) and ridge (Hoerl and Kennard 1970) penalties. The LASSO penalty promotes sparsity and performs variable selection through shrinkage: the lowest regression coefficients, corresponding to the least informative predictors, are attributed a zero value. The ridge penalty accommodates correlated variables and ensures numerical stability. The calibration of the tuning parameters, the overall penalty and mixing proportion for the two penalties were determined by minimizing the prediction root mean squared error (RMSE) using 10-fold cross-validation (i.e. the data were partitioned into 10 subsets; for each of these subsets, the data were trained on the other 9 partitions and fitted on the given left-out subset over which the RMSE was estimated). To prevent over-fitting, the optimal calibration parameters were defined as those providing the most sparse model (as measured by the number of non-zero regression coefficients), among those yielding an RMSE within one standard error of the minimum RMSE (Meinshausen and Bühlmann 2006).

Sparse partial least squares regression

Partial least squares regression is a supervised dimension reduction technique that builds summary variables as linear combinations of the original set of variables. To ensure that the resulting lower-dimension representation of the data is relevant to the outcome of interest, the components are defined iteratively such that they explain as much of the remaining covariance between the predictors and the (health) outcome as possible. The sPLS approach simultaneously yields good predictive performance and appropriate variable selection by creating sparse linear combinations of the original predictors (Chun and Keleş 2010). Sparsity is induced by including a penalty (η) in the estimation of the linear combination coefficients, i.e. all coefficients with an absolute value lower than some fraction η of the maximum absolute coefficient are shrunk to zero. This procedure is called soft thresholding (Lenters et al. 2014). Only the first K components are included as covariates in a linear regression model. The values of K and η were calibrated by minimizing the RMSE using 5-fold cross-validation (the default implementation). To complete model comparison, we generalized the reference implementation such that it also includes the empty model (K=0).

Graphical Unit Evolutionary Stochastic Search

As part of the Bayesian variable selection approaches, GUESS seeks for models that optimally predict the health outcome. Each model is defined by a unique combination of covariates (Bottolo and Richardsony 2010). Method estimation calls upon the identification of the most relevant models among the 2^p (where p denotes the total number of covariates) possible combinations of covariates using an evolutionary Monte Carlo algorithm, which combines tempered multiple chains run together with genetic algorithms. These ensure both improved mixing of the sampler and exchange of information across chains (Bottolo et al. 2013).

For each simulated data set, we ran the GUESS algorithm for 20,000 iterations and discarded the first 5,000 to account for burn-in. We set the number of chains to 3. To ease convergence and prevent extensive parameter calibration, noting E the a priori expected model size and ρ its variance, we set $E = 3$ and $\rho = 3$ for k<5, and $E = k + 2$ and $\rho = 5$ for $k \geq 5$. As a conservative measure, among the models visited we retained those associated with a posterior probability above 0.01.

From the union of all exposures included in these models retained, we selected those with a marginal posterior probability of inclusion (MPPI; the probability that a variable is included in any of the models retained) greater than the (1-0.05/237) quantile of the MPPI distribution under the null hypothesis (i.e. where no covariate was associated to the outcome).

The original goal of GUESS is to select the best combination(s) of covariates to predict the outcome. Its latest implementation (Liquet et al. 2015) allows posterior simulation of the coefficients estimates for a given model. However, in our simulation context where the true predictors are different from one dataset to the other, this indirect (i.e. conditional on variable selection) estimation procedure would require integrating posteriors over all models visited, which represents a prohibitive computational effort and is therefore incompatible with a direct coefficient estimation. As a conservative alternative, we used an additional ridge regression step with the variables selected by GUESS to estimate the methods' coefficients. This procedure is however likely to lower the quality of the estimates.

Deletion-Substitution-Addition algorithm

DSA is an iterative linear regression model search algorithm (Sinisi and van der Laan 2004). The set of potential models is limited by three user-specified constraints: the maximum order of interaction amongst predictors, the maximum power for a given predictor and the maximum model size. At each iteration, the following three steps are allowed: 1) removing a term, 2) replacing one term with another, and 3) adding a term to the current model. The search for the best model starts with the intercept model and identifies an optimal model for each model size.

The final model is selected by minimizing the value of the RMSE using 5-fold cross-validated data. We allowed no polynomial or interaction terms, and considered models including up to 40 covariates (this number was never reached in our simulations).

We used R implementations of the statistical methods under investigation, which are available in the packages *stats*, *glmnet, spls, R2GUESS* and *DSA*, respectively. The R codes developed by the authors and the correlation matrix Σ are provided in Supplemental Material S2, and Supplemental Material, Excel File Table S1, respectively.

Statistical Performance Assessment

The performances of each statistical method were evaluated using key criteria measuring the relevance of the variable selection and the quality of the point estimates.

The sensitivity of a method was calculated for each scenario and simulation as the proportion of true predictors that were actually selected by the given method. The specificity was calculated the same way as the proportion of unrelated exposures that were not selected.

The false discovery proportion (FDP) was defined as the proportion of selected variables that were not genuinely related to the outcome. When no variable was selected in a given run, we considered no variable was mistakenly selected and the FDP was given a value of 0%. FDP was not computed for scenarios with 0 true predictors.

We investigated the accuracy of the estimated coefficients by means of the mean absolute bias calculated over the 237 coefficient estimates as

$$\frac{1}{237}\sum_{i=1}^{237}|\beta_i - \hat{\beta}_i|$$

where β_i represents the coefficient used in the simulation, and $\hat{\beta}_i$ the corresponding estimate. The mean absolute bias was also computed over the true predictors and over the unrelated exposures (i.e. non true predictors) separately.

Owing to the possibly strong correlations between exposures, the argument could be made that not selecting a true predictor but instead picking up another highly correlated variable, should not be seen as a complete false selection, in the sense that the statistical method did not fully missed the signal. In order to account for this in our study, we defined alternative sensitivity and FDP measures accounting for such a partial agreement, based on the highest absolute correlation estimated between the true predictors and the covariates selected by the statistical method:

$$AltSens = \frac{1}{k}\sum_{i\in A} \underset{j\in B}{max}\{|\widehat{corr}(X_i, X_j)|\}, AltFDP = 1 - \frac{1}{n_B}\sum_{j\in B} \underset{i\in A}{max}\{|\widehat{corr}(X_i, X_j)|\},$$

A is the set of true predictors and B the set of variables selected by the method (also called hits), k and $n_B$being their respective sizes. $AltSens$ measures the average highest absolute correlation value between a true predictor and any variable selected by the method; $AltFDP$ measures the average highest absolute correlation value between a selected variable and any of the true predictors. If the set of selected covariates includes all true predictors, these alternative metrics correspond to the classical sensitivity and FDP measures. Given that $|\widehat{corr}(X_i, X_j)| \leq 1$, $AltSens$ is always greater than sensitivity, and $AltFDP$ is always smaller than FDP.

Extended variable selection protocol

The argument could be made that in order to increase sensitivity and avoid missing important signals, one should not look at the selected exposures only but also consider all exposures highly correlated (i.e. at a level greater than α, α varying between 0.6 and 0.9) to these hits. Resulting sensitivity and FDP was computed for this approach.

RESULTS

Correlation structure used for generating exposures

The $\mathbf{\Sigma}$ matrix is defined as the nearest positive definite matrix to the INMA correlation structure, and only marginally differed from its parent: 75% of the absolute differences were smaller than 0.01 and 95% were smaller than 0.05. The large majority (83%) of absolute correlations between exposures in Σ were lower than 0.2, but 78% of the exposures were correlated at a level >0.6 with at least one other exposure (Figure S1).

Performance assessment for scenarios set 1

The simulation results of scenarios set 1 are presented in Figures 1 and 2 and Table 1.With true predictors drawn fully at random, the per-scenario average (standard error) absolute pairwise correlation amongst true predictors ranged between 0.12 and 0.15 (0.12 and 0.16).

Over all investigated numbers of true predictors (i.e. k=0,1,2,3,5,10,25), the EWAS approach yielded a sensitivity greater than 90%, but a specificity as low as 46% and a FDP greater than 67% (due to the selection of a large number of exposures as measured by n_B/k in Table 1). The alternative FDP ranged between 24% and 45% across simulations. The mean absolute bias was large (range, 0.02 to 0.47), but restricted to the true predictors only, it was the smallest of all statistical methods (≤0.10 vs. ≥0.30 for all other methods, Figure S2).

When EWAS was followed by a multiple linear regression step (EWAS-MLR), the FDP improved over all scenarios (range, 30% to 80%), as well as the specificity (> 95% over all scenarios), at the cost however of a much lower sensitivity (<56% over all scenarios). The alternative sensitivity was between 38% and 87%, while the alternative FDP was between 16% and 34%. The mean absolute bias was large (9.00 on average over all scenarios).

Results were similar while using other corrections for multiple testing (Figure S3). If no adjustment for multiple comparison was applied, the FDP obtained with this modified EWAS was greater than 89% and $AltFDP$ was greater than 42%.

GUESS, sPLS, ENET and DSA methods all showed lower FDP than EWAS or EWAS-MLR. On average (5th percentile; 95th percentile) over all scenarios and on these four statistical methods, sensitivity was 78% (60% ; 91%), FDP was 39% (21% ; 62%) specificity was 96% (89% ; 100%), alternative sensitivity was 95% (91% ; 99%) while alternative FDP was 12% (5% ; 20%). The mean absolute bias was 0.03 (0.00 ; 0.11), and 0.52 (0.32 ; 0.89) when restricted to the true predictors only (Figure S2). These methods selected on average 1.79 times the number of true predictors (n_B/k in Table 1). On average, DSA and GUESS proved a better compromise between sensitivity and FDR (average values: 80% and 33% respectively) than sPLS and ENET (average values: 78% and 44%, respectively), with DSA slightly favoring a high sensitivity while GUESS favored a low FDP (Figure 2). Yet, none of these statistical methods outperformed the others across all scenarios and indicators investigated.

Over all methods, as the number of true predictors increased, the variable selection performances generally decreased: FDP and $AltFDP$ substantially increased across all statistical methods (on average, +29%, +9% between k=1 and k=25, respectively), sensitivity and AltSens slightly decreased for all methods but EWAS-MLR and ENET (-7% and -4% between k=1 and k=25, respectively), and mean absolute bias increased (especially for the EWAS-based approaches). Sensitivity and AltSens largely decreased for EWAS-MLR, and largely increased for ENET. However, care should be taken in interpreting these trends since an increased number of true predictors is accompanied by an increased signal to noise ratio (R^2 of the true model), but also by an increased risk that some true predictors are highly correlated.

Performance assessment under alternative versions of the scenarios

Scenarios in which true exposures were selected so that all their absolute pairwise correlations were <0.2 (set 2) or >0.5 (set 3) showed that the higher the level of correlation amongst the true predictors, the lower the sensitivity for the ENET, GUESS and DSA methods (and to a lower extent for the EWAS-MLR method); and the higher the mean absolute bias, mostly for the EWAS-based and DSA approaches (Figure S4). FDP was impacted for the ENET, sPLS and DSA methods, although not in a consistent direction. Apart from a large $AltFDP$ decrease for the ENET method, the specificity and the alternative definitions of both sensitivity and FDP were poorly impacted. Note that selecting predictors with high pairwise correlation yielded an increase in the variance of the error term used in the simulations.

Generating exposures from a correlation matrix with higher (scenarios set 4) or lower (scenarios set 5) levels of correlation (Figure 3) did not alter the methods' comparison, but had a major impact on the sensitivity, FDP and mean absolute bias: the higher the correlation among the exposures, the worse the performances of the methods. With correlation levels divided by two compared to scenarios set 1, the sensitivity was greater than 85% for all scenarios and statistical methods (except ENET for k<3) and FDP decreased on average by 23% compared to the same scenarioin set 1. The alternative sensitivity and FDP and the specificity were less sensitive to the overall correlation of exposures, and less consistently affected.

Deviation from the assumption of normally distributed exposures (scenarios set 6) led to analogous results compared to scenarios set 1, except for EWAS-MLR method showing better results for the bootstrapped data, yet not competing with the other methods (Figure S5).

Considering varying effect sizes for true predictors (drawn from a uniform distribution in [0.5,1.5], scenarios set 7) did not alter the methods comparison and had a limited impact on the statistical performances: sensitivity and AltSens were moderately lower (-10% to -7% on average compared to same scenario, set 1), and specificity, mean absolute bias (except for EWAS-based methods), FDP and AltFDP were not impacted (Figure S6).

Extended variable selection protocol

In scenarios set 1, when augmenting the list of variables selected by a method with variables that were correlated to any these hits above some threshold α, a substantial increase in FDP was observed (except for EWAS-based methods for which FDP was already high), even for α as high as 0.8 or 0.9 (Table S2).

DISCUSSION

We tested the ability of several established statistical approaches to identify, from a large set of correlated exposures, those causally related to a continuous health outcome. We mostly relied on sensitivity and false detection proportion to assess the statistical methods' performances: specificity was always high in our simulations (which can be at least partially attributed to our assumption that no more than 25 of the 237 exposure variables were associated with the outcome) making FDP a more discriminating criterion. In addition to the classical measures of sensitivity and FDP, we introduced alternative definitions that account for the fact that false positives that are correlated to a true predictor might actually provide information that can be used to identify this true predictor.

The EWAS-related approaches performed poorly under the scenarios investigated. EWAS captured a large number of (falsely positive) covariates (average FDP for scenarios set 1, 86%), irrespective of the procedure used for correcting multiple hypothesis testing (Benjamini and Hochberg, Benjamini and Yekutieli and permutation-based FDR procedures, or Bonferroni correction). This is mostly due to FDR procedures assuming the statistics (i.e. here, the p-values) are unbiased, while in our simulations there was a high potential for confounding due to independently fitting regression models on correlated exposures. However, compared to the other methods investigated, EWAS best estimated the true predictors coefficients values. When EWAS was followed by a multiple linear regression step (EWAS-MLR), a small proportion of true predictors were captured (average sensitivity for scenarios set 1, 33%). Yet, these two statistical methods still performed much better than if no correction for multiple comparisons was applied, which in the literature corresponds to the association of each exposure with the outcome being considered sequentially in different publications. For these two methods, the alternative FDP remained relatively high (32% on average for scenarios set 1), suggesting that in the investigated scenarios, many of the variables selected by these approaches were not strongly correlated to a true predictor.

Using the ENET, sPLS, GUESS and DSA approaches, most true predictors were selected by the method (average sensitivity of 78% for scenarios set 1) and a substantial proportion of exposures were mistakenly suspected to be associated with the outcome (average FDP of 39% for scenarios set 1). For these four statistical methods, exposures that were mistakenly selected were on average highly correlated to at least one of the true predictors (average $AltFDP$ of 12% for scenarios set 1). Similarly, when a true predictor was not selected by these methods, it was likely that a highly correlated covariate was selected instead (average $AltSens$ of 95% for scenarios set 1). None of the multivariate statistical methods tested clearly outperformed the others across all scenarios and properties examined. Globally, DSA and GUESS proved the best compromise between sensitivity and FDP, with DSA favoring highest sensitivity and GUESS favoring lowest FDP. Deviating from the assumption of normally distributed exposures or from the assumption of even effect sizes for true predictors did not alter the methods comparison. However, GUESS and DSA were the most affected by high correlation levels amongst true predictors (scenarios set 3), whereas EWAS and sPLS were less sensitive to this feature. Other factors such as ease of use, ability to force in confounders, accommodation for different study designs (e.g. longitudinal designs) or for non-linear exposure-response relations (e.g. using splines) may also be important for choosing between these methods.

The argument has been made that selecting variables highly correlated to the true predictors should not be considered as a false selection per-se (Frommlet et al. 2012), and our alternative definitions of FDP and sensitivity were actually developed under this logic. As indicated by the relatively high values of these modified criteria for the four multivariate statistical methods, most of the true predictors are likely to belong to the set of exposures highly correlated to the variables selected. Considering the "hits" and their correlated covariates may thus be a way to capture the true predictors. There are several things to note when considering such an extended variable selection protocol and our findings in general: (i) in genetic studies, one can identify known and unknown correlated polymorphisms by utilizing the architecture of the genome; this may not apply to the exposome as correlations between exposures may arise from a variety of mechanisms (diet, social economic status, etc.), and there is no guarantee that selecting a correlated variable will provide useful information on the causal mechanism linking the true predictors to the outcome. As such, the distinction between true predictors and predictors correlated to those true predictors is challenging; (ii) lowering the threshold for selection (by including all predictors correlated to a selected predictor) will likely lead to an increased FDP under the usual definition, which may more than offset the benefits (in terms of an increased sensitivity). This is exemplified in our results for this protocol which suggested a substantial increase of the FDP when selecting variables correlated at a level greater than 0.8 with the hits (Table S2). In that respect, it is important to stress that our alternative definition of FDP ($AltFDP$) is not the FDP that would result from the variable selection method induced by $AltSens$ where predictors highly correlated to the selected ones would be additionally selected. Instead, it is the FDP that would result from using the original selection protocol, but counting correlated variables as "true predictors", with a weight proportional to their correlation with the true predictor.

Our simulation work extends that of Sun et al. (2013) to a more realistic context for the exposome in terms of number of exposures and of their correlation structure. We showed that the correlation structure under which the exposures are generated greatly impacts the performances of the statistical methods (Figure 3), meaning that the results from Sun et al. and of any simulation study with fixed correlation structure cannot be generalized in a straightforward way to the exposome context.

Our study relied on several modeling assumptions which need to be taken into consideration while discussing the generality of our results. First, we assumed no effect measure modification of a covariate on the health outcome by any other covariate (departure from additivity), a situation which may in practice not be true. Incorporating interactions terms would strongly increase the size of the modeling space (e.g. in this study, 27966 first order interactions) and would require extending our statistical methods to test for interactions, using dedicated techniques from all families investigated here (e.g. Lie and Zhang 2005). Withdrawing the restriction of binary effect sizes and incorporating varying effect sizes in the simulation did not alter the FDP; and it only artificially reduced the statistical power to detect weaker effects (reduction in sensitivity of 10% on average). This can be explained by a ceiling effect, i.e. the already high sensitivity could not be improved for exposures with higher than average effects to the same extent as it could be reduced for exposures with lower than average effects. Overall, the induced sensitivity loss was consistent across all methods and did not help in further discriminating the statistical methods under investigation. Importantly, we did not consider measurement error or misclassification in exposure covariates, while these have a potentially large impact on statistical power and bias, in particular in the case of classical type error (de Klerk et al. 1989; Rappaport et al. 1995, Perrier et al. in press). As a result, method performances may be hampered in real-life situations, but there is no a priori reason to think that statistical methods under investigation in this study would be differentially affected by these issues. We further assumed that exposures were normally distributed. Deviating from this assumption did not alter the performances of the methods. Finally, similarly to Sun et al. (2013), we used a limited set of statistical methods all borrowing from the linear regression framework. Alternative approaches such as profile regression, cluster analysis or other machine learning methods could complement this portfolio of approaches but could not be straightforwardly compared with our set of regression-based approaches.

CONCLUSIONS

Relying on a realistic exposome structure, we screened a large set of correlated exposures out of which only a small number were directly associated with a continuous outcome. Our results suggests that the multivariate methods investigated should be preferred to univariate approaches to investigate the exposome: despite not achieving a low FDP, they show satisfactory statistical performances and represent different balances between sensitivity and FDP. Based on our performance metrics, we identified DSA and GUESS as providing somehow better performances, but this was not true across all scenarios and properties examined, and in real case analyses, methodological choices should also be guided by computational complexity and flexibility considerations such as the possibility to accommodate for confounders. Performances of the statistical methods were strongly influenced by the correlation among the exposome covariates, illustrating an issue inherent to the exposome research, namely that the statistical methods investigated are not able to efficiently differentiate between true predictors and correlated covariates.

References

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. . Ser. B 57: 289–300.

Benjamini Y, Yekutieli D. 2001. The control of the False Discovery rate in multiple testing under dependency. Ann. Stat. 29: 1165–1188.

Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità [in Italian]. Pubbl. del R Ist. Super. di Sci. Econ. e Commer. di Firenze.

Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, Newcombe P, et al. 2013. GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm. PLoS Genet. 9: e1003657.

Bottolo L, Richardsony S. 2010. Evolutionary stochastic search for bayesian model exploration. Bayesian Anal. 5: 583–618.

Buck Louis GM, Yeung E, Sundaram R, Laughon SK, Zhang C. 2013. The exposome - Exciting opportunities for discoveries in reproductive and perinatal epidemiology. Paediatr. Perinat. Epidemiol. 27: 229–236.

Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al. 2013. Deciphering the Complex : Methodological Overview of Statistical Models to Derive OMICS-Based Biomarkers. Environ. Mol. Mutagen. 54:542–557; doi:10.1002/em.

Chun H, Keleş S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72: 3–25.

De Klerk NH, English DR, Armstrong BK. 1989. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. Int. J. Epidemiol. 18: 705–712.

Frommlet F, Ruhaltinger F, Twaróg P, Bogdan M. 2012. Modified versions of Bayesian Information Criterion for genome-wide association studies. Comput. Stat. Data Anal. 56: 1038–1051.

Greenland S. 1994. Hierarchical regression for epidemiologic analyses of multiple exposures. Environ. Health Perspect. 102:33–39; doi:10.1289/ehp.94102s833.

Guxens M, Ballester F, Espada M, Fernández M, Grimalt J, Ibarluzea J, et al. 2012. Cohort Profile: The INMA –INfancia y Medio Ambiente–(Environment and Childhood) Project. Int. J. Epidemiol. 930–940.

Higham NJ. 2002. Computing the nearest correlation matrix - A problem from finance. IMA J. Numer. Anal. 22: 329–343.

Hoerl AE, Kennard RW. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12: 55–67.

Lenters V, Portengen L, Smit L, Jönsson B, Giwercman A, Rylander L, et al. 2014. Phthalates, perfluoroalkyl acids, metals and organochlorines and reproductive function: a multipollutant assessment in Greenlandic, Polish and Ukrainian men. Occup. Environ. Med. 1–9; doi:10.1136/oemed-2014-102264.

Lenters V, Portengen L, Rignell-Hydbom A, Jönsson B, Lindh C, Piersma A, et al. 2015. Prenatal Phthalate, Perfluoroalkyl Acid, and Organochlorine Exposures and Term Birth Weight in Three Birth Cohorts: Multi-Pollutant Models Based on Elastic Net Regression. *Environ Health Perspect*; DOI:10.1289/ehp.1408933.

Li, F., Zhang N R. 2010. Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics. J. Am. Stat. Ass.105:1202-1214.

Liquet B, Bottolo L, Campanella G, Richardson S, Chadeau-Hyam M. 2015. In press. R2GUESS: a Graphics Processing Unit-Based R Package for Bayesian Variable Selection Regression of Multivariate Responses. J. Stat. Softw.

Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. Ann. Stat. 34: 1436–1462.

Patel CJ, Bhattacharya J, Butte AJ. 2010. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One 5: e10746.

Patel CJ, Ioannidis JP a. 2014. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. J. Epidemiol. Community Health 1–5; doi:10.1136/jech-2014-204195.

Patel CJ, Rehkopf DH, Leppert JT, Bortz WM, Cullen MR, Chertow GM, et al. 2013. Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States National Health and Nutrition Examination Survey. Int. J. Epidemiol. 42:1795–810; doi:10.1093/ije/dyt208.

Perrier F, Giorgis Allemand L, Slama R, Philippat C. In press.Within-subject pooling of biological samples as a way to reduce exposure misclassification in biomarker-based studies of chemicals with high temporal variability. Epidemiology.

Rappaport SM. 2011. Implications of the exposome for exposure science. J. Expo. Sci. Environ. Epidemiol. 21:5–9; doi:10.1038/jes.2010.50.

Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. 2014. The blood exposome and its role in discovering causes of disease. Environ. Health Perspect. 122: 769–774.

Rappaport SM, Symanski E, Yager JW, Kupper LL. 1995. The relationship between environmental monitoring and biological markers in exposure assessment. Environ. Health Perspect. 103: 49–53.

Sinisi SE, van der Laan MJ. 2004. Loss-Based Cross-Validated Deletion/Substitution/Addition Algorithms in Estimation. U.C. Berkeley Div. Biostat. Work. Pap. Ser. 3.

Slama R, Vrijheid M. 2015. Some Challenges of Studies Aiming to Relate the Exposome to Human Health. Occup. Environ. Med.; doi:10.1136/oemed-2014-102546.

Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, et al. 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ. Heal. 12:85; doi:10.1186/1476-069X-12-85.

Tibshirani R. 1996. Regression Selection and Shrinkage via the Lasso. J. R. Stat. Soc. B 58: 267–288.

Tzoulaki I, Patel CJ, Okamura T, Chan Q, Brown IJ, Miura K, et al. 2012. A nutrient-wide association study on blood pressure. Circulation 126:2456–64; doi:10.1161/CIRCULATIONAHA.112.114058.

Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. 2014. The Human Early-Life Exposome (HELIX): Project Rationale and Design. Environ. Health Perspect. 122:535–544; doi:10.1289/ehp.1307204.

Wild CP. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol. biomarkers Prev. 14:1847–50; doi:10.1158/1055-9965.EPI-05-0456.

Wild CP. 2012. The exposome: from concept to utility. Int. J. Epidemiol. 41:24–32; doi:10.1093/ije/dyr236.

Willett WC. 2002. Balancing life-style and genomics research for disease prevention. Science 296: 695–698.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67: 301–320.

Table 1. Statistical performances of the statistical methods for scenarios set 1. Results are given as mean [5th; 95th] percentiles over all scenarios (100 runs per scenario).

Method	**Sensitivity**	**AltSens**	**FDP**	**AltFDP**	**Specificity**	n_B/k	**Mean absolute bias**	**Mean absolute bias for TP**
EWAS	0.96 [0.84;1.00]	0.97 [0.93;1.00]	0.86 [0.75;0.98]	0.37 [0.27;0.59]	0.72 [0.18;1.00]	11.27 [0.00;40.02]	0.59 [0.02;2.12]	0.04 [0.00;0.16]
EWAS-MLR	0.33 [0.00;1.00]	0.59 [0.16;1.00]	0.58 [0.00;1.00]	0.27 [0.00;0.73]	0.99 [0.94;1.00]	0.86 [0.00;3.00]	9.00 [0.00;69.39]	0.67 [0.00;1.00]
ENET	0.66 [0.00;1.00]	0.92 [0.19;1.00]	0.37 [0.00;1.00]	0.11 [0.00;0.61]	0.97 [0.94;1.00]	1.15 [0.00;2.60]	0.02 [0.00;52.40]	0.74 [0.00;1.00]
sPLS	0.86 [0.80;1.00]	0.96 [0.87;1.00]	0.52 [0.50;0.97]	0.16 [0.08;0.51]	0.90 [0.25;1.00]	3.59 [0.00;29.52]	0.03 [0.02;2.12]	0.46 [0.00;0.20]
GUESS	0.88 [0.00;1.00]	0.97 [0.25;1.00]	0.39 [0.00;1.00]	0.10 [0.00;0.52]	0.98 [0.93;1.00]	1.45 [0.00;2.20]	0.02 [0.00;26.35]	0.37 [0.00;1.00]
DSA	0.73 [0.70;1.00]	0.94 [0.82;1.00]	0.28 [0.50;0.96]	0.09 [0.12;0.46]	0.99 [0.33;1.00]	0.95 [0.00;22.46]	0.04 [0.02;2.12]	0.51 [0.00;0.30]

AltFDP: Alternative definition of the false discovery proportion (see methods section for definition); AltSens: Alternative definition of the sensitivity (see methods section for definition); DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; FDP: False Discovery Proportion; GUESS: Graphical Unit Evolutionary Stochastic Search; n_B/k: number of variables selected by the method (n_B) over the number of true predictors (k); sPLS: Sparse partial least-squares; TP: True Predictors.

Figures Legends

Figure 1. Performances of the statistical methods for scenarios set 1. Model performances are summarized by their sensitivity (A), alternative sensitivity ($AltSens$, see methods section) (B), false detection proportion (FDP) (C), alternative FDP ($AltFDP$, see method section) (D), specificity (E) and mean absolute bias (F). For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot) and the variability of each statistics is summarized using the one standard error both ways from the average value (vertical dotted line). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Figure 2. Sensitivity and FDP for scenarios set 1. For each scenario defined by a number of true predictors varying from 0 to 25, for each statistical method, sensitivity and FDP over the 100 runs are summarized by their mean value. DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Figure 3. Performances of the statistical methods according to the amount of correlation between the exposures. Model performances are summarized by their sensitivity (A), alternative sensitivity ($AltSens$, see methods section) (B), false detection proportion (FDP) (C), alternative FDP ($AltFDP$, see method section) (D), specificity (E) and mean absolute bias (F). The full line connects results for exposures generated from a multivariate normal distribution with covariance matrix Σ (scenarios set 1); the dashed line with covariance matrix Σ^- (correlations divided by two compared to Σ, scenarios set 4) and the dotted with covariance matrix Σ^+ (correlations multiplied by two compared to Σ and upper bounded by 1, scenarios set 5). For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Figure 1.

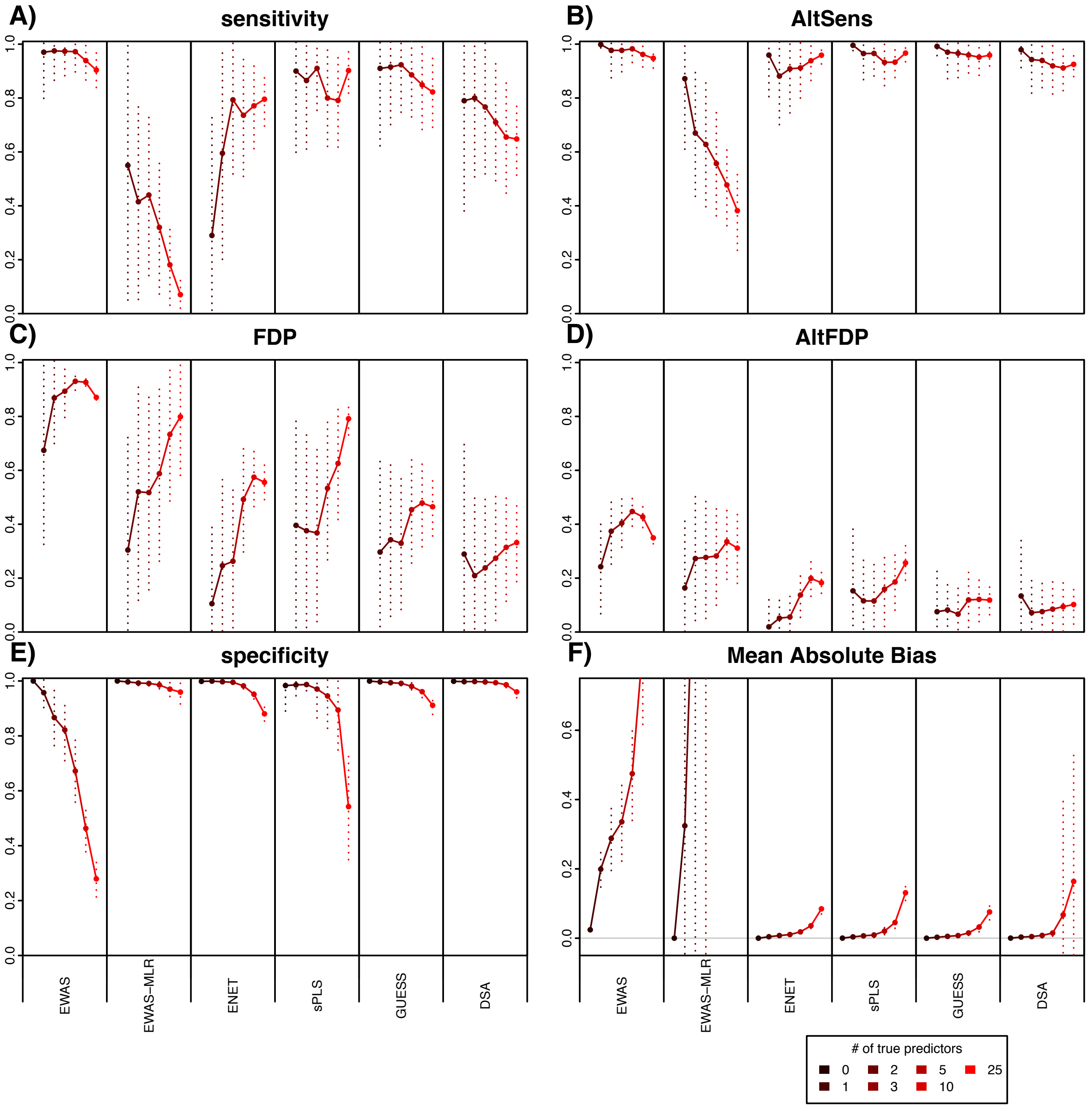

A)
sensitivity
B)
AltSens
C)
FDP
D)
AltFDP
E)
specificity
F)
Mean Absolute Bias
0.0
0.2
0.4
0.6
0.8
1.0
EWAS
EWAS−MLR
ENET
sPLS
GUESS
DSA
# of true predictors
0
1
2
3
5
10
25

Figure 2.

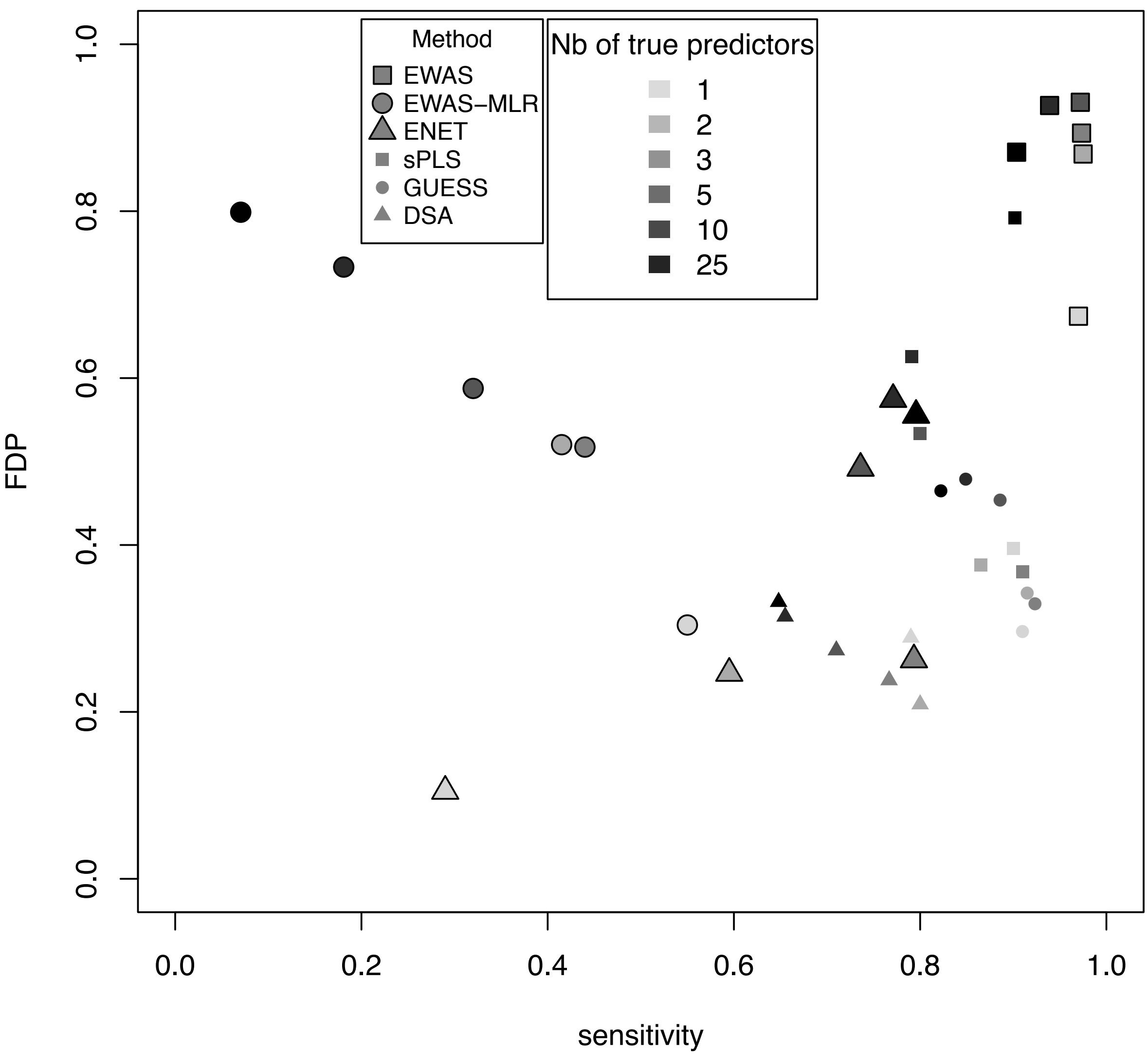
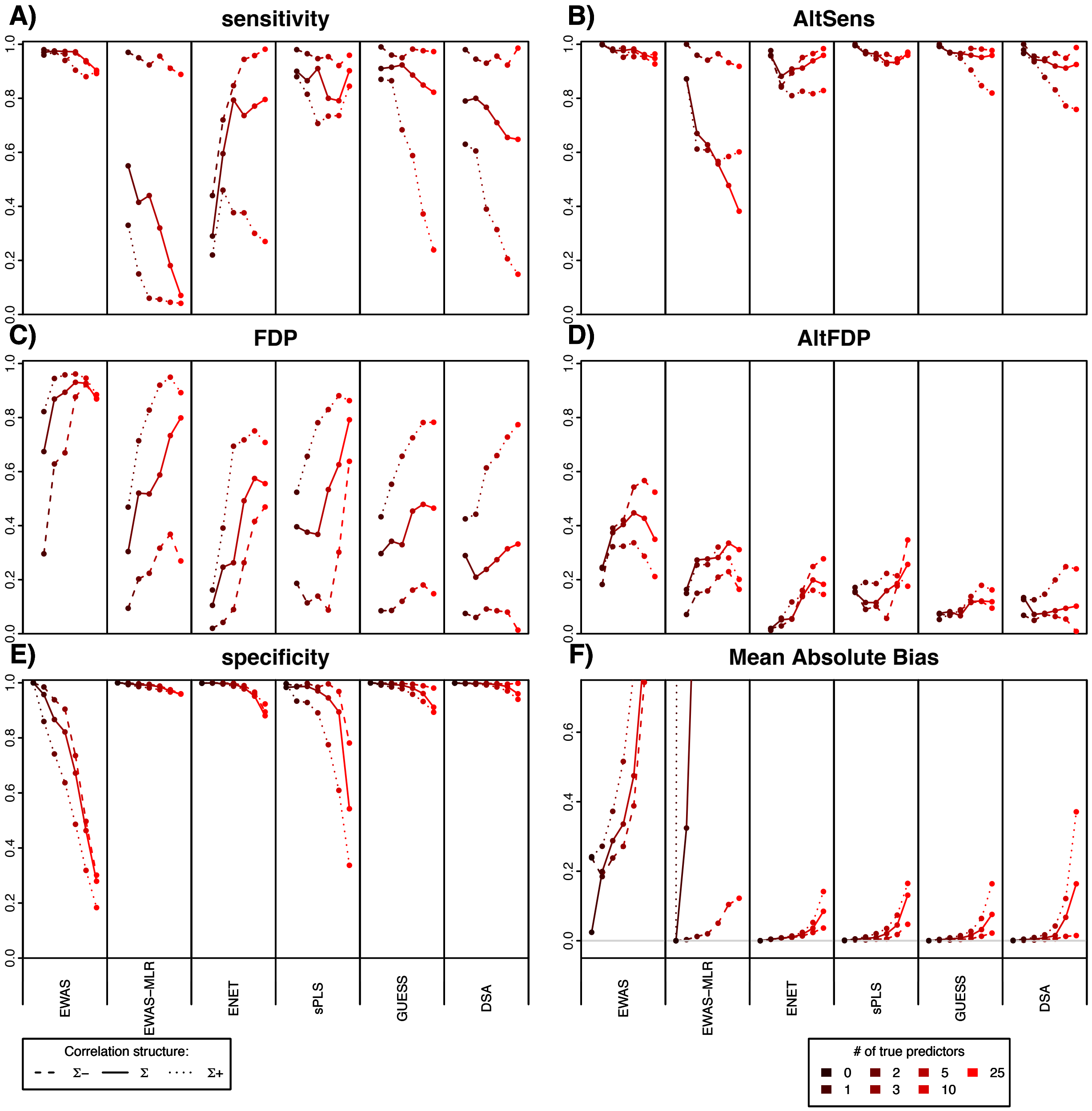

Method
EWAS
EWAS−MLR
ENET
sPLS
GUESS
DSA
Nb of true predictors
1
2
3
5
10
25
FDP
0.0
0.2
0.4
0.6
0.8
1.0
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0

Figure 3.


A)
sensitivity
B)
AltSens
C)
FDP
D)
AltFDP
E)
specificity
F)
Mean Absolute Bias
0.0
0.2
0.4
0.6
0.8
1.0
EWAS
EWAS–MLR
ENET
sPLS
GUESS
DSA
Correlation structure:
Σ−
Σ
Σ+
# of true predictors
0
1
2
3
5
10
25